20

TM-5243/005/00

# INTERACTIVE SYSTEMS RESEARCH: INTERIM REPORT TO THE DIRECTOR, ADVANCED RESEARCH PROJECTS AGENCY, FOR THE PERIOD 16 SEPTEMBER 1975 to 15 MARCH 1976





15 APRIL 1976



THIS REPORT WAS PRODUCED BY SDC IN PERFORMANCE OF CONTRACT DAHC 15-73-C-0080, ARPA ORDER NO. 2254, PROGRAM CODE NO. 6P10.

SYSTEM DEVELOPMENT CORPORATION
2500 COLORADO AVENUE • SANTA MONICA, CALIF. 90406

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle) (6) Interactive Systems Research: Interim Report to the Director, Advanced Research Projects Agency, for the Period 16 September 1975 to 15 March 1976 | | 5. TYPE OF REPORT & PERIOD COVERED Technical--9/75-3/76 |
| | | 6. PERFORMING ORG. REPORT NUMBER (14) SDC-TM-5243/005/00 |
| 7. AUTHOR(s) Bernstein, M. I. / Bernstein (10) | | 8. CONTRACT OR GRANT NUMBER(s) (15) DAHC15-73-C-0080 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS System Development Corporation 2500 Colorado Avenue Santa Monica, California 90406 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order-2254 Program Code No. 6P10 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209 | | 12. REPORT DATE (11) 15 Apr[redacted]76 |
| | | 13. NUMBER OF PAGES i, 16 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (12) 20p. | | 15. SECURITY CLASS. (of this report) Unclassified |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Cleared for public release; distribution unlimited

(9) Technical rept. Sep 75-Mar 76,

D D C

MAY 11 1977

C

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

speech-understanding systems
acoustic phonetics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)


Error-analysis experiments were conducted on four acoustic-phonetic-analysis programs being developed for a speech-understanding system. The experiments were conducted primarily to identify areas of parametric or signal-processing error that may be corrected in a later version of the system.


DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

339 9..

mit

TM-5243/005/00

# INTERACTIVE SYSTEMS RESEARCH: INTERIM REPORT TO THE DIRECTOR, ADVANCED RESEARCH PROJECTS AGENCY, FOR THE PERIOD 16 SEPTEMBER 1975 to 15 MARCH 1976




AUTHOR: M. I. BERNSTEIN

829-7511, EXT. 2086


15 APRIL 1976


THIS REPORT WAS PRODUCED BY SDC IN PERFORMANCE OF CONTRACT DAHC 15-73-C-0080, ARPA ORDER NO. 2254, PROGRAM CODE NO. 6P10.





**System Development Corporation**

2500 Colorado Avenue • Santa Monica, California 90406

## TABLE OF CONTENTS



## LIST OF TABLES

## 1. INTRODUCTION

This Interim Report covers System Development Corporation's (SDC's) Speech Understanding Research (SUR) activities during the six months from September, 1975, to March, 1976. At the beginning of that period, advanced versions of SDC's SUR system components had been implemented and were being tested in preparation for a year-end demonstration to the ARPA SUR Steering Committee. During the period, the testing was completed and the demonstration was held. Since the demonstration (in January), several of the system components have been modified and expanded, a new control and language-processing component has been installed, and comprehensive performance testing has continued. By the end of the current year, we will have completed the construction, testing, and demonstration of a prototype speech-understanding system that operates in accordance with the specifications set forth in 1971 by an ARPA Study Group chaired by Allen Newell. The Study Group summarized the specifications as follows:

> The system should: accept continuous speech from many cooperative speakers of the general American dialect, in a quiet room, over a good quality microphone, allowing slight tuning of the system per speaker, but requiring only natural adaptation by the user, permitting a slightly selected vocabulary of 1,000 words, with a highly artificial syntax, and a task like the data management [task], with a simple psychological model of the user, providing graceful interaction, tolerating less than 10% semantic error, in a few times real time, and be demonstrable in 1976 with a moderate chance of success.

SDC has progressed toward meeting these specifications by defining, implementing, and evaluating successively more powerful and more refined versions of a small 1971 prototype that grew out of a predecessor Voice Input/Output Project. That original prototype was, in turn, based on a system that had been built by P. J. Vicens and D. R. Reddy at Stanford University.

During the 1972-1973 contract year, two prototype systems were constructed: one that contained a sophisticated control and linquistic "top end," or parser, and limited acoustic-phonetic support, and one that had limited linguistic support but relatively complete acoustic-phonetic algorithms for all of the

phoneme classes. During the following year, the two systems were combined in a system that had strong capabilities in both linguistic and acoustic-phonetic processing. At the same time, additional components were developed to store and access the system's growing lexicon and to serve as sources of knowledge to aid understanding in a specific task environment. During 1974 and 1975, all of these components were expanded and refined in the effort that led to the system demonstrated in January.

A significant part of SDC's current research effort has been directed at defining the acoustic-phonetic properties of speech in algorithmic form and determining which properties are most useful in speech understanding by computer.

This interim report summarizes the results of error analyses conducted of four principal acoustic-phonetic programs: those for labeling vowels and sonorants, fricatives and plosives, and syllable segments, and for measuring acoustic stress.

## 2. PROGRESS IN ACOUSTIC PHONETICS

Progress in acoustic-phonetics includes detailed error analyses of the vowel/sonorant, fricative/plosive, syllable-segmentation, and acoustic-stress-measurement programs.

### 2.1 ERROR ANALYSIS OF VOWEL/SONORANT PROGRAM

The purpose of the vowel/sonorant analysis program (VOWSON) is to locate, define the boundaries of, and appropriately label steady-state, voiced, non-fricated areas of a speech utterance. A complete description of the program is given elsewhere [1,2].

An experiment was conducted to determine where VOWSON fails; knowing that will give us a better understanding of parametric errors (and perhaps errors in the signal-processing techniques used to derive the parameters), of necessary design

modifications, and of the coarticulation process. Although all these problems cannot be completely resolved, it is necessary to provide the mappers with the most reliable information possible regarding the types of confusions that are likely to occur, their probability, and their phonetic environments. The mappers can then operate well even though the VOWSON parameter set is faulty. As the program is improved, the information it provides to the mappers will reflect the improvements.

Four male subjects were selected for the experiment. All four spoke some form of general American dialect, but there were some differences between them. Each subject spoke 21 sentences. Although the 21 sentences were designed primarily to fulfill syntactic and semantic needs, an analysis of the transcribed sounds indicated they also contained a phonetic balance close to that generally found in English when compared to the studies of Shoup [3] and Denes [4].

All of the recording was done in the SDC SUR laboratory,which has a signal-to-noise ratio of approximately 50 dB. Each subject was prompted on a Tektronix interactive terminal, which showed him the next sentence to say. The sentences were spoken into a high-quality microphone. The resulting signal was digitized at the rate of 20,000 samples per second and stored on tape. A Raytheon 704 computer was used for all of the data acquisition and analysis.

The 84 recorded sentences were transcribed by a phonetician (Peter Ladefoged), who listened to them under ideal conditions in a sound-proofed room. He also examined spectrograms and computer-produced acoustic analyses before completing his transcriptions. The transcriptions used a machine representation for transcribing English, as shown in Table 1.

Stress was marked in accordance with the analysis given by Ladefoged [5] so that a consistent distinction was made among stressed, unstressed, and reduced vowels. Each vowel or consonant symbol was assigned to a given time period in the utterance. The phonetician attempted to do this in a consistent way

TABLE 1. ARPABET REPRESENTATION OF VOWEL/SONORANT PHONEMES

| Phoneme | ARPABET Representation | Phoneme | ARPABET Representation |
|---|---|---|---|
| i | IY | u | UW |
| ɪ | IH | ə | AX |
| ɛ | EH | ɨ | IX |
| æ | AE | ɝ | ER |
| a | AA | r | R |
| ʌ | AH | w | W |
| ɔ | AO | l | L |
| o | OW | m | M |
| ʊ | UH | n | N |
| y | Y | ŋ | NX |

(for instance, by assigning the beginning of the Y in "The U.S." to the point in time where the second formant was at a maximum), but many segment boundaries were assigned only arbitrarily (for instance, the division between Y and UW in "The U.S." was made without any explicit algorithm).

The transcription differed from a conventional phonetic transcription in that an attempt was made to label all the vowel qualities that could be heard, subject to the limitation that only the ARPABET character code could be used. Thus, if the vowel in the word "length" was pronounced as a glide with two distinct parts--AE as in "had" and IY as in "heed"--the word was transcribed as L AE IY NX TH. However, in short unstressed vowels in which the quality could not be precisely determined by ear, AX (the schwa vowel, as in the first syllable of "about") was transcribed. In many of these cases, the acoustic analyses showed that the vowel in fact contained well-defined formants with some other interpretable quality; in these cases, it may well be proper to consider the phonetic transcription to be wrong.

Table 2 shows the overall statistics in terms of the number of vowels and sonorants labeled and unlabeled.

TABLE 2. OVERALL STATISTICS

| | RG | WB | BR | RW | Total |
|---|---|---|---|---|---|
| Total vowels | 202 | 194 | 191 | 194 | 781 |
| Unlabeled vowels | 7 | 15 | 32 | 17 | 71 |
| Total sonorants | 103 | 111 | 112 | 110 | 436 |
| Unlabeled sonorants | 8 | 16 | 36 | 18 | 78 |

A total of 1217 sounds were used as data in this analysis. An additional 104 sounds that were transcribed as Y or diphthongs were omitted. (Diphthongs were not included because by definition they involved patterns of formant movement. Y was omitted because at present there is no unique steady-state formant pattern for Y other than IY, and its recognition probably involves the relative heights of $F_2$ and $F_3$ at their maximum within the segment.) Twelve percent, or 149, of the 1217 sounds were unlabeled, and approximately half of these were due to speaker BR, whose vowels (particularly those in unstressed syllables) tended to be very short. Ninety-two of the 1068 labeled sounds were broken into multiple segments unaccounted for by the transcription; most of these cases should be regarded as being due to the insufficiency of the transcription, rather than being considered errors of the program. Approximately 6% of the 1217 sounds occurred in areas where there had been formant-tracking errors; almost 80% of these occurrences were in nasal areas.

There were only six occurrences in which a sound was not labeled vowel or sonorant but passed onto the fricative-plosive segmenter and labeler. There were 16 occurrences in which the same segment was labeled both vowel/sonorant and fricative/plosive.

In assessing the accuracy of the segmentation, we must distinguish between errors due to a faulty location of a sound and errors due to an incorrect determination of the number of sounds. Of the total 1217 sounds, we can properly assess the accuracy of the location of the boundaries of 976. Of the remainder, 149, or 12%, were left unsegmented or unlabeled, and 92, or 7%, were segmented into more sounds than were transcribed. Table 3 compares the hand transcriptions and the program segmentations. (It should be remembered that the transcriptions often very arbitrarily assigned temporal boundaries and cannot be considered as reliable as the program.)

TABLE 3. DEGREE OF AGREEMENT IN DURATION BETWEEN HAND TRANSCRIPTIONS AND MACHINE SEGMENTATIONS

| Percentage of 976 sounds within: | | | |
|---|---|---|---|
| 20 msec. | 30 msec. | 40 msec. | 50 msec. |
| 71 | 83 | 91 | 95 |

Table 4 is a confusion matrix of first-choice labels for all vowels and sonorants. The transcribed sounds are shown in the rows, and the machine-assigned labels are shown in the columns. The sounds M, N, and NX are considered to be correctly labeled when labeled NA. The column heading "OTH" refers to sounds labeled fricative and/or plosive, and the column heading "MIS" refers to the sounds that were not segmented or labeled (i.e., the label is missing).

Most of the sounds, with the exception of AX, were confused in a non-random way with other sounds. By detailed analysis, the confusions can be examined for patterns. For example, the vowel IY was labeled correctly 89 out of 125 occurrences. In four cases there were no labels; in 32 cases there was an incorrect first-choice label. On close examination, we find that most of the errors fall into one or the other of two groups. In the first group, there

TABLE 4. CONFUSION MATRIX OF FIRST-CHOICE LABELS

| | IY | IH | EH | AE | AA | AH | AO | OW | UH | UW | AX | IX | ER | R | W | L | NA | OTH | MIS | TOT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IY | 89 | 9 | | | | | | | | 9 | | 5 | 1 | | 4 | 1 | 3 | | 4 | 125 |
| IH | 3 | 45 | 4 | | | | | | 10 | 4 | 4 | 25 | 4 | | 2 | 5 | 4 | | 19 | 129 |
| EH | 1 | 1 | 10 | 8 | | 5 | | | 1 | 1 | 1 | 1 | | 2 | | | 1 | | 1 | 33 |
| AE | | | 2 | 32 | 9 | 8 | | | 1 | 3 | 1 | | | 1 | | 1 | | | 1 | 56 |
| AA | | | | 1 | 19 | | 5 | 2 | 5 | 1 | 1 | | | | | | | | 3 | 37 |
| AH | | | | 3 | 13 | 19 | | 1 | 3 | 1 | 7 | 1 | | 1 | | | 1 | | 1 | 51 |
| AO | | | | | 1 | 1 | 2 | 2 | | | 1 | | 2 | | | | | 1 | 5 | 15 |
| OW | | | | | | | | 14 | 3 | | 2 | | | | 2 | 1 | | | 1 | 23 |
| UH | | | | | 1 | | 1 | 2 | 1 | 1 | | | | | | 1 | | | | 7 |
| UW | 1 | | | | | | | | 1 | 19 | | 6 | | | 2 | 1 | | | | 30 |
| AX | | | 2 | | 3 | 2 | | 7 | 30 | 14 | 19 | 11 | 3 | 2 | 1 | 4 | 5 | 1 | 25 | 129 |
| IX | 2 | 1 | | | | | | | 4 | 3 | | 64 | 1 | | | 1 | | | 3 | 79 |
| ER | | | 1 | | | | | 2 | | 4 | | 3 | 30 | 14 | | 1 | 4 | | 8 | 67 |
| R | | | | | | | | 1 | | | 1 | | 8 | 6 | | | 2 | 2 | 13 | 33 |
| W | | | | | | | | | 2 | 4 | | | | | 13 | 2 | 2 | | 11 | 34 |
| L | | | | | | | 2 | 13 | 3 | 8 | 1 | 1 | 1 | | 9 | 50 | | 2 | 22 | 112 |
| M | 3 | | | | 1 | | | | | 8 | 1 | | | | | 3 | 91 | | 9 | 116 |
| N | | 2 | 2 | | | | | | | 17 | 2 | 6 | | | 1 | 6 | 74 | 1 | 25 | 136 |
| NX | | | | | | | | | | | | | | | 1 | | 4 | | | 5 |

are 17 occurrences of labels having a combination of the choices NA, L, W, UW. These occur in a total of three words: submar<u>ines</u> (11 cases), submar<u>ine</u> (2 cases), and man<u>y</u> (4 cases); they always occur in a nasal environment, and in all cases the other contiguous phone is a sonorant. Moreover, eight of these occurrences are in areas in which formant-tracking errors occurred (i.e., the formant tracker tracked the wrong peaks as $F_1$, $F_2$, or $F_3$). It is likely that in the other cases, either proper peaks were not picked or the LPC model was inadequate in nasal areas.

Another group of 14 errors all had label combinations of IX, IH, IY (i.e., if IY was present it was not the first choice). This confusion is more acceptable than the first group of errors because the sounds IX and IH are closer to IY. This second group of errors occurred in a larger set of words: "many," "submarine," "Lafayette," "Seawolf," "me," "nuclear," "the," "torpedo." However, all but five were in a nasal or sonorant environment. The one error in the 32 incorrect labels that did not fall into either group was found in the word "nuclear"; the IY had been labeled an ER.

## 2.2 ERROR ANALYSIS OF FRICATIVE/PLOSIVE PROGRAM

The fricative/plosive program performs a segmentation and labeling of voiced and unvoiced fricatives and plosives in a continuous speech utterance. Narrow-window linear prediction spectra are used to extract parameters, and time sequences of the parameters are then used for segmentation. Phonemic labels are assigned to the derived segments on the basis of the dynamic features of the parameters. A complete description of the program has been given earlier [6,7,8].

A program evaluation experiment was undertaken employing the same 84-sentence corpus (21 sentences X 4 male subjects) used to evaluate the vowel-sonorant analysis program. The digitized speech input, speech analysis programs, and phonetic transcription are identical to those described elsewhere. With regard to the fricative and plosive portions of the utterances, the phonetician noted that his transcription is not always reliable; in a number of cases the sound found by the computer may be a more accurate record of the phonetic event than the sound transcribed. This is particularly probable in the case of voiced-voiceless confusions, where, as the phonetician pointed out, he was occasionally influenced by phonemic considerations despite his best intentions.

General labeling performance is summarized in Table 5. Less than 3 percent of non-fricative and non-plosive segments were erroneously labeled as fricative or plosive; most of these errors (28 of 37) are nasals mislabeled as /v/ or /ð/. The program is thus quite reliable in segmenting fricatives and plosives from the continuous speech utterances that are its input.

Labeling performance is shown in a detailed confusion-matrix format in Table 6. Results for the four speakers have been pooled. Several points are important in interpreting the confusion matrix. In order to present results in this format given the three-choice, associated-score format produced by the program, we defined a relative score threshold. The first-choice label, plus all others scoring within 18 points of the first-choice label, are counted as labels for the transcribed segment. This approach was taken to properly evaluate the

TABLE 5. MISLABELING OF SEGMENTS

| Transcribed Segment | Total Segments Transcribed | Total Fricative and Plosive Labels | Percent of Transcribed |
|---|---|---|---|
| m | 116 | 19 | 16 |
| n | 135 | 11 | 8 |
| l | 113 | 2 | 2 |
| y | 25 | 3 | 12 |
| æ | 55 | 2 | 4 |
| all other vowels and sonorants | 877 | 0 | 0 |
| TOTALS | 1321 | 37 | 3 |

TABLE 6. SUMMARY OF LABELING PERFORMANCE AND CONFUSION MATRIX (SEE TEXT)

| TRANSCRIBED SEGMENTS | | LABELS DETERMINED BY FRICATIVE-PLOSIVE PROGRAM | | | | | | | | | | | | | | | | | | OMITTED SEGMENTS (see note) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Percent of Transcribed | | | Confusion Matrix | | | | | | | | | | | | | | | | |
| LABEL | TOTAL | CORRECT | FALSE | OMITTED | ʃ | s | z | f | θ | h | p | t | k | g | d | ɾ | b | ð | v | VS | NF |
| ʃ | 23 | 83 | 26 | 4 | 19 | 4 | | | | | | 1 | 1 | | | | | | | | 1 |
| s | 204 | 94 | 13 | 0 | 3 | 191 | 12 | 4 | 4 | | 1 | | | | | | | | 2 | | |
| z | 51 | 73 | 57 | 2 | | 15 | 37 | 2 | 2 | 1 | | | | | | | 1 | 4 | 4 | 1 | |
| f | 40 | 88 | 100 | 3 | | | | 35 | 34 | | | | | | | | | 3 | 3 | | 1 |
| θ | 13 | 69 | 92 | 0 | | | 1 | 9 | 9 | | | | 1 | | | | | 1 | | | |
| h | 82 | 52 | 7 | 45 | | | | 2 | 2 | 43 | 2 | | | | | | | | | 14 | 23 |
| p | 28 | 46 | 100 | 25 | | 1 | | | | 1 | 13 | 4 | 2 | 2 | | | 10 | 8 | | | 7 |
| t | 84 | 52 | 63 | 36 | | 1 | | 3 | 3 | 5 | 20 | 44 | 6 | 5 | 13 | | 3 | 2 | 1 | | 30 |
| k | 36 | 57 | 53 | 22 | | 1 | | 2 | 2 | 3 | 2 | 3 | 21 | 2 | | | | 4 | | 1 | 7 |
| g | 16 | 56 | 156 | 6 | 2 | | | | | | | | 6 | 9 | 5 | | 4 | 4 | 4 | | 1 |
| d | 110 | 33 | 66 | 24 | | | 1 | 1 | | 1 | 9 | 5 | 1 | 4 | 36 | 3 | 26 | 24 | 17 | 13 | 13 |
| ɾ | 34 | 44 | 97 | 62 | | | | | | | | | | | | 15 | 16 | 17 | | 13 | 8 |
| b | 72 | 42 | 122 | 36 | | | | | | 1 | 5 | | | 2 | 11 | 11 | 30 | 37 | 21 | 12 | 14 |
| ð | 67 | 57 | 94 | 33 | | | 1 | 5 | 5 | 1 | 7 | | 2 | 2 | 4 | 3 | 15 | 38 | 18 | 12 | 10 |
| v | 60 | 45 | 85 | 43 | | | | 8 | 7 | 1 | | | 2 | 1 | | | 7 | 25 | 27 | 7 | 19 |

NOTE: **Omitted segments (ones not labeled by the fricative-plosive program) are split into two categories: VS contains those which were erroneously labeled as vowel and/or sonorant by another program[6]; NF contains segments not found by either program.**

scoring algorithms of the program. However, a side effect of allowing multiple labels for one transcribed segment is that the total number of labels found in a transcribed-phoneme row in Table 6 exceeds the number of segments transcribed for that row. Likewise, the percentages shown total more than 100 percent, since the "correct" and "false" columns are the percentages of transcribed segments represented by the diagonal and off-diagonal labels, respectively.

Twenty rows of the full confusion matrix, corresponding to the transcribed phonemes /i, I, ɛ, æ, a, ʌ, ɔ, ʊ, U, y, u, ə, ɨ, ɝ, r, w, l, m, n, ŋ/, have been omitted from Table 6 for clarity; the few mislabelings of these phonemes are listed in Table 5. It should further be noted that the program presently does not attempt to assign the label /ʒ/.

## 2.3 ERROR ANALYSIS OF SYLLABLE SEGMENTATION PROGRAM

A syllable segmentation program is used to automatically subdivide the incoming speech waveform into syllabic units. To make segmentation decisions, the program applies a convex hull function to a sonorant energy function. A technical description of the program is given in [ 1].

The program was evaluated using the same four-speaker, 21-sentence (84-utterance) data base used for testing the vowel/sonorant and fricative/plosive programs. All 84 utterances were manually segmented and labeled by a phonetician. Both phoneme-segment boundaries and syllable-segment boundaries were marked. The program was considered to have found the correct syllable boundary if the machine-generated boundary differed by not more than 20 msec. from the manually transcribed boundary. Otherwise, the boundary was considered incorrect. Additional boundaries were also considered in the evaluation; these were the cases in which the program found a syllable boundary not transcribed by the phonetician. Missed boundaries were also taken into account in the scoring; these are the cases in which the program failed to mark a boundary that was manually transcribed. The results are presented in Table 7.

TABLE 7. SUMMARY OF SYLLABLE SEGMENTATION PROGRAM ANALYSIS

| Speaker | Transcribed Boundaries | Wrong Boundaries | Extra Boundaries | Missed Boundaries |
|---|---|---|---|---|
| WB | 192 | 19 | 6 | 15 |
| BR | 194 | 11 | 8 | 15 |
| RW | 201 | 17 | 9 | 13 |
| RG | 193 | 29 | 13 | 7 |
| Totals | 780 | 76 | 36 | 50 |

## 2.4 ERROR ANALYSIS OF ACOUSTIC STRESS MEASUREMENT PROGRAM

In the mapping strategy, it is important to know the acoustic stress of each syllable. There are three reasons for this. First, reduced vowels (primarily schwa) are distinguished more by their stress level than by their formant frequency structure. Second, in a "bottom-driving" strategy (in which words are located and recognized purely on the basis of acoustic clues), it is important to begin the bottom-driving with a stressed syllable, since this will contain more reliable acoustic-phonetic information than a syllable with a lower stress level. Third, agreement between predicted stress levels and machine-generated stress levels is a part of the scoring function of the mapper.

Acoustic stress is assigned to each syllable on the basis of three parameters: duration, intensity, and relative pitch. These parameters and the manner in which they are used to assign stress are discussed in detail elsewhere [ 1]. The program assigns four levels of stress: stressed (S), medium stressed (M), non-stressed (N), and reduced (R).

An experiment was conducted to determine how well these automatically-assigned labels compared with labels perceived and manually transcribed by a trained phonetician. Four stress levels were manually assigned to each syllable in the same 84-sentence test corpus described above. It would have been preferable

to manually assign only three levels (stressed (1), unstressed (2), reduced (3)), since it was perceptually difficult for the transcriber to distinguish among more than three distinct stress levels. For this reason, the results of the machine processing were combined so that the machine-assigned stress had to be more than one stress level different from the perceived stress to register as an error. Therefore, machine-generated stress was considered correct if it agreed with perceived stress as in the following table.

| Perceived Stress | Acceptable Machine-Assigned Stress |
|---|---|
| 1 | S, M |
| 2 | S, M, N |
| 3 | M, N, R |
| 4 | N, R |

Table 8 summarizes the results of the experiment. It indicates the number of vowels of each of the four stresses perceived by the transcriber and the number (and percentage) assigned to each stress level by the program. The number of errors tended to increase with the number of syllables in an utterance. Some errors occurred because of a syllable-boundary omission. This led, for example, to a perceived reduced syllable being called stressed; the duration of the syllable was very long, and this acted to increase the machine-generated stress level (since duration is one of the parameters used to assign stress). Additional syllable boundaries had the opposite effect, i.e., the machine-generated stress was lower than it should have been.

**TABLE 8. RESULTS OF ERROR ANALYSIS OF ACOUSTIC STRESS MEASUREMENT PROGRAM**

| Speaker | Stress 1 | | | Stress 2 | | | Stress 3 | | | Stress 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Number Perceived | Number Assigned S or M | % | Number Perceived | Number Assigned S, M, or N | % | Number Perceived | Number Assigned M, N, or R | % | Number Perceived | Number Assigned N or R | % |
| RW | 89 | 62 | 69.66 | 8 | 8 | 100 | 58 | 52 | 90 | 66 | 53 | 80.30 |
| WB | 80 | 63 | 78.75 | 10 | 10 | 100 | 40 | 37 | 92.5 | 88 | 65 | 73.86 |
| BR | 83 | 72 | 86.75 | 10 | 9 | 90 | 50 | 44 | 88 | 76 | 49 | 64.47 |
| RG | 51 | 45 | 88.24 | 36 | 34 | 94.44 | 38 | 32 | 84.21 | 95 | 70 | 73.68 |
| Overall | 303 | 242 | 79.87 | 64 | 61 | 95.31 | 186 | 165 | 88.71 | 325 | 237 | 72.92 |

## 3. OTHER ACCOMPLISHMENTS

The complete Milestone System, with a 600-word vocabulary and the parser developed by Stanford Research Institute (SRI) was demonstrated and described to the ARPA SUR Steering Committee at the end of January. The only major component that was not completed in time for the demonstration was the Lexical Subsetter, a program that deletes unlikely candidates from a list of proposed words to be mapped, thus saving considerable computation time. The demonstration was relatively successful in that, of the several utterances chosen and spoken by a member of the Steering Committee, one was successfully processed, although the processing time required for all the utterances was considerably greater than desirable.

Subsequent to the demonstration to the Steering Committee, the Lexical Subsetter was completed, tested, and integrated with the other components of the Lexical Analyzer. Simultaneously, a test driver developed by SDC to test the Lexical Analyzer in isolation was expanded to become a complete parser to replace the SRI parser. The new SDC parser supports all of the interrogative statements in the SRI parser, but does not support ellipsis or imperative and declarative statements. It also incorporates a lookaside memory to eliminate redundant mapping. The major advantage of the new version of the system is that it reduces the processing time (exclusive of the acoustic-phonetic processing) by more than a factor of ten. This will permit much more thorough testing of the total system while consuming significantly less computer resources.

In preparation for completion and testing of the five-year system, the 1,000-word lexicon and associated additions to the grammar for the new parser were selected. The control strategy has been modified to incorporate a set of variables that can be set to change priorities of processing choices, such as depth- vs. breadth-first parsing. This has been done to permit more thorough evaluation and understanding of the contribution of individual components and their interactions under varying control strategies.

Preliminary testing of the system has highlighted where effort should be concentrated to improve the system to obtain maximum performance. It has also clearly demonstrated that a "missed" utterance requires significantly more processing time than one that is correctly understood. We are examining several methods to terminate "unprofitable" processing without diminishing the understanding level.

## 4. PLANS

The project will proceed with the final implementation and testing of the Speech Understanding System, although with some changes in emphasis. The major change in emphasis will be the suspension of work by SRI on their parser (the system's "top end") and substitution of a simpler version that retains a majority of the required functions, but requires significantly less computer resources to operate and test. From April through June, work will be concentrated on expanding the lexicon from 600 to 1,000 words and on removing the known deficiencies from the Acoustic-Phonetic Processor, the Lexical Analyzer, and other components; reducing the total time (and resources) required to process an utterance; and preparing a plan, the necessary tools, and system modifications needed to perform testing and evaluation of the system. From July through September, the system will be tested according to the proposed test plan. The results will be analyzed, and a demonstration of the Five-year System will be made. The final report, which will be published in November, will present detailed results of performance testing and evaluation.

## 5. REFERENCES

[1] Bernstein, M. I., Interactive Systems Research: Final Report to the Director, Advanced Research Projects Agency for the Period 16 September 1974 to 15 September 1975. SDC TM-5243/004/00, 15 November 1975.

[2] Kameny, Iris, "Automatic Acoustic-Phonetic Analysis of Vowels and Sonorants," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, April 1976.

[3] Shoup, June, "Phoneme Selection for Studies in Automatic Speech Recognition," JASA 34(4), April 1962.

[4] Denes, P. B., "On the Statistics of Spoken English," JASA 35(6), June 1963.

[5] Ladefoged, Peter, A Course in Phonetics. New York: Harcourt, Brace, Jovanovich, 1975.

[6] Molho, L. M., "Automatic Recognition of Fricatives and Plosives in Continuous Speech Using a Linear Prediction Method," JASA 55:411 (abstract), 1974.

[7] Molho, L. M., "Automatic Recognition of Fricatives and Plosives in Continuous Speech," IEEE Symposium on Speech Recognition, IEEE Catalog No. CH0878-9AE, 68-73 (1974).

[8] Molho, Lee, "Automatic Acoustic-Phonetic Analysis of Fricatives and Plosives," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, April 1976.